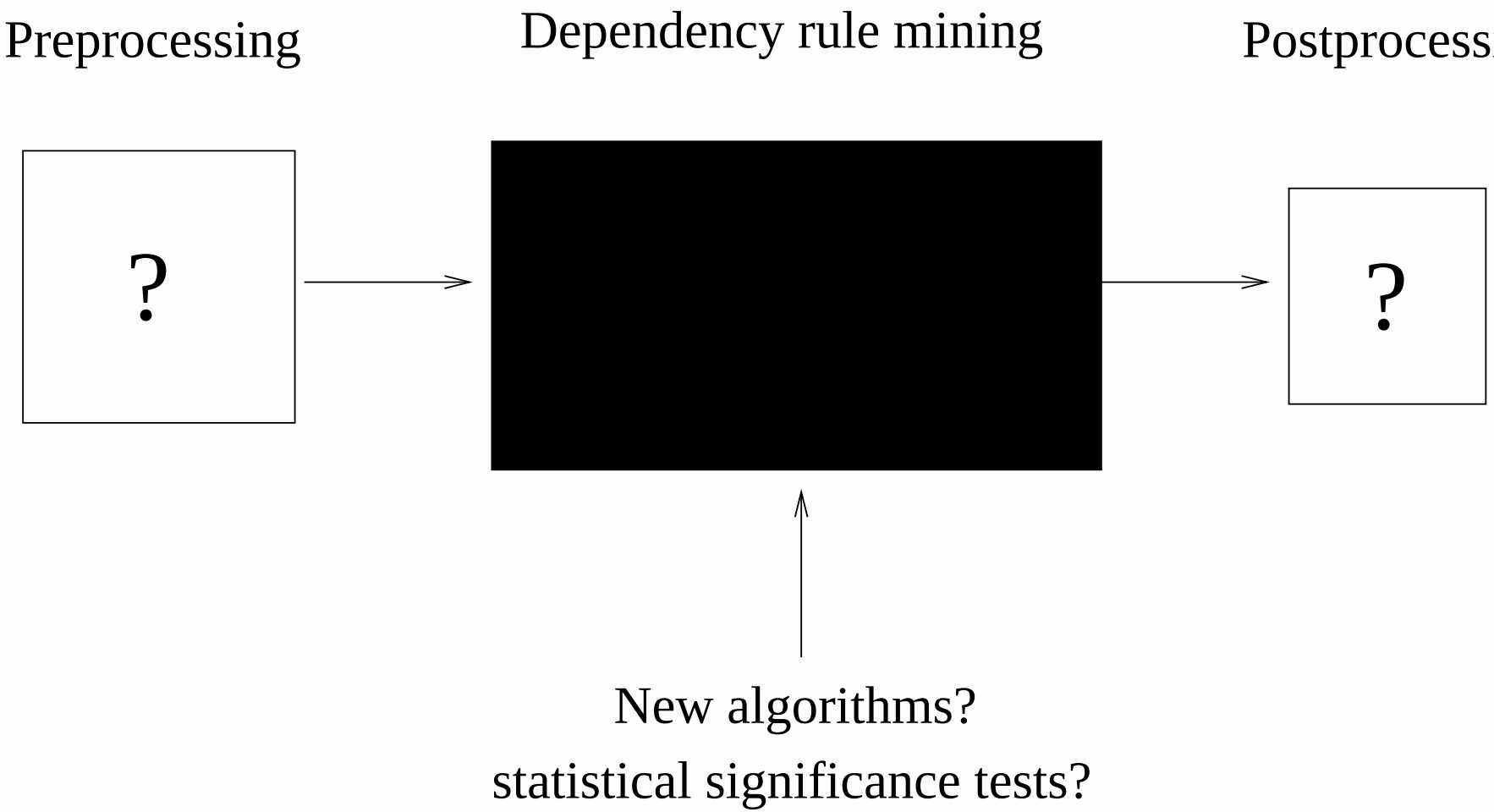
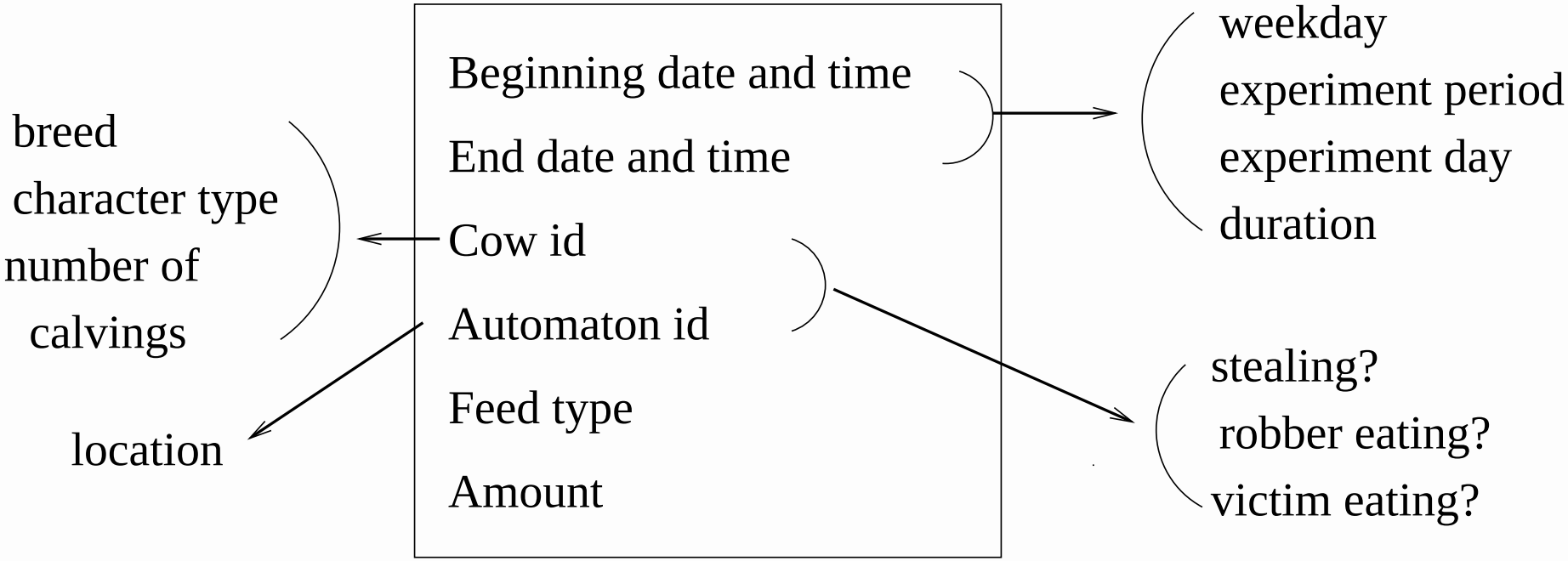
Thorough analysis of log data with dependency rules: Practical solutions and theoretical challenges



Typical log data

- timestamps (beginning, end)
- user id
- device id
- event type
- objects
- other characteristics

Example: Log from cows' feeding automata



Dependency rule $X \rightarrow A$

- expresses statistical dependency
- evaluated by statistical goodness measures (e.g. p, χ^2, MI)
- no minimum frequency thresholds
- ideally non-redundant

  Given “Rainy day $\rightarrow$ Wet dog”,
  “Rainy day and Autumn $\rightarrow$ Wet dog”
  hardly adds any new information.

Dependency rules are attractive but restricted

Efficient algorithms & globally optimal results

but should be binary data!

How to handle log data?

Main problems in log data

1. How to balance between groups and individuals? (common trends, individual regularities and peculiarities)
2. How to discretize numerical and periodic variables optimally?
3. How to extract attributes from intrinsic dimensions? (spatial & temporal contexts, interactions)

1. Balancing between groups and individuals

1.1 Incorporating attribute hierarchies

- background information, different abstraction levels

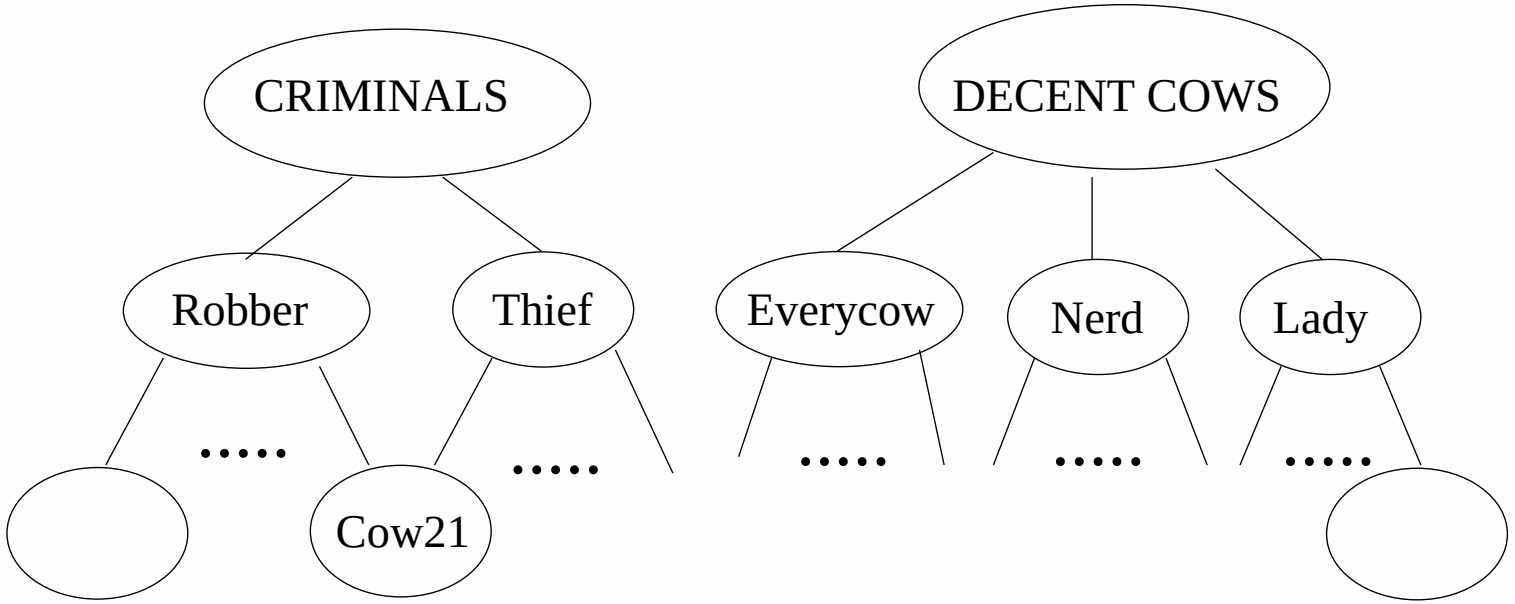
1.2 Handling individual differences in numerical variables

- New features which describe normality or exceptionality of values!
  (w.r.t. individual or reference group)

1.3 Focusing on individuals and exceptional events

- Search for $X \rightarrow A$ or $QA \rightarrow B$, where A interesting

1.1 Incorporating attribute hierarchies



- How to compare $thief \wedge Q \rightarrow A$ to $cow21 \wedge Q \rightarrow A$? or $(thief \vee robber) \wedge Q \rightarrow A$?
- Challenge: Statistical test + search algorithm!
- Small hierarchies can be handled with extra attributes and constraints

2. Discretizing numerical and periodic variables

- Problem: Different discretizations of A can be optimal for different dependencies!
- $\Rightarrow$ Challenge: How to discretize dynamically during the search?
- Preprocessing tricks can simulate dynamic discretization
  - works when the number of numerical variables is small
  - otherwise use coarser atomic intervals or concentrate on exceptional values

Example (tricks + Kingfisher)

rule	$\ln(p)$	fr
$P1day < 5, Amount_{NI} < 0.0 \rightarrow Theft$	-1005	1343
$P1day < 6, Amount_{NI} < 0.0 \rightarrow Theft$	-942	1382
$P1day < 5, Amount_{NI} < 0.5 \rightarrow Theft$	-882	1343
$P1day < 5, Amount_{NI} < 1.0 \rightarrow Theft$	-813	1343
$P1day < 5, Amount_{NI} < 1.5 \rightarrow Theft$	-781	1343
$P1day < 6, Amount_{NI} < 0.5 \rightarrow Theft$	-763	1382
$P1day < 5, Amount_{NI} < 2.0 \rightarrow Theft$	-763	1343

$P1day$ = day from the beginning of period 1
$Amount_{NI}$ = individually normalized amount

3. Opening intrinsic dimensions

A new research area with great potential!

- Spatial context
  - What is happening in the neighbouring devices?
- Temporal context
  - Event history of an individual, a device, or the whole cowhouse?
- Social interactions

Conclusions

- Practical tricks suffice with “simple” log data
- But log data can be complex!
  - log data from web-stores (detailed market baskets + user profiles)
  - mobile devices with GPS and sensors
  - combining log data from different devices
- Important problems in analyzing other heterogeneous real world data, as well!